全国赛培训系列讲座：

关于模型的认识、理解与实现

Based on MATLAB & Python

出品人

讲师介绍

@王小川_MATLAB

王小川，博士，金融领域从业者，终生学习者。关注神经网络、数据挖掘、统计分析应用领域，国内最大的MATLAB论坛管理员，曾多次参与Mathworks公司培训活动，近年在北京、上海、武汉等地举办多次数据分析与挖掘研讨会与培训，有丰富的实战技巧与培训经验，其微博上的发布的数据挖掘公开课程总点击量超过50万。

目录
contents

1

First Part

分析流程与思路

2

Second Part

数据分析软件介绍

3

Third Part

实战演练

第一部分

First Part

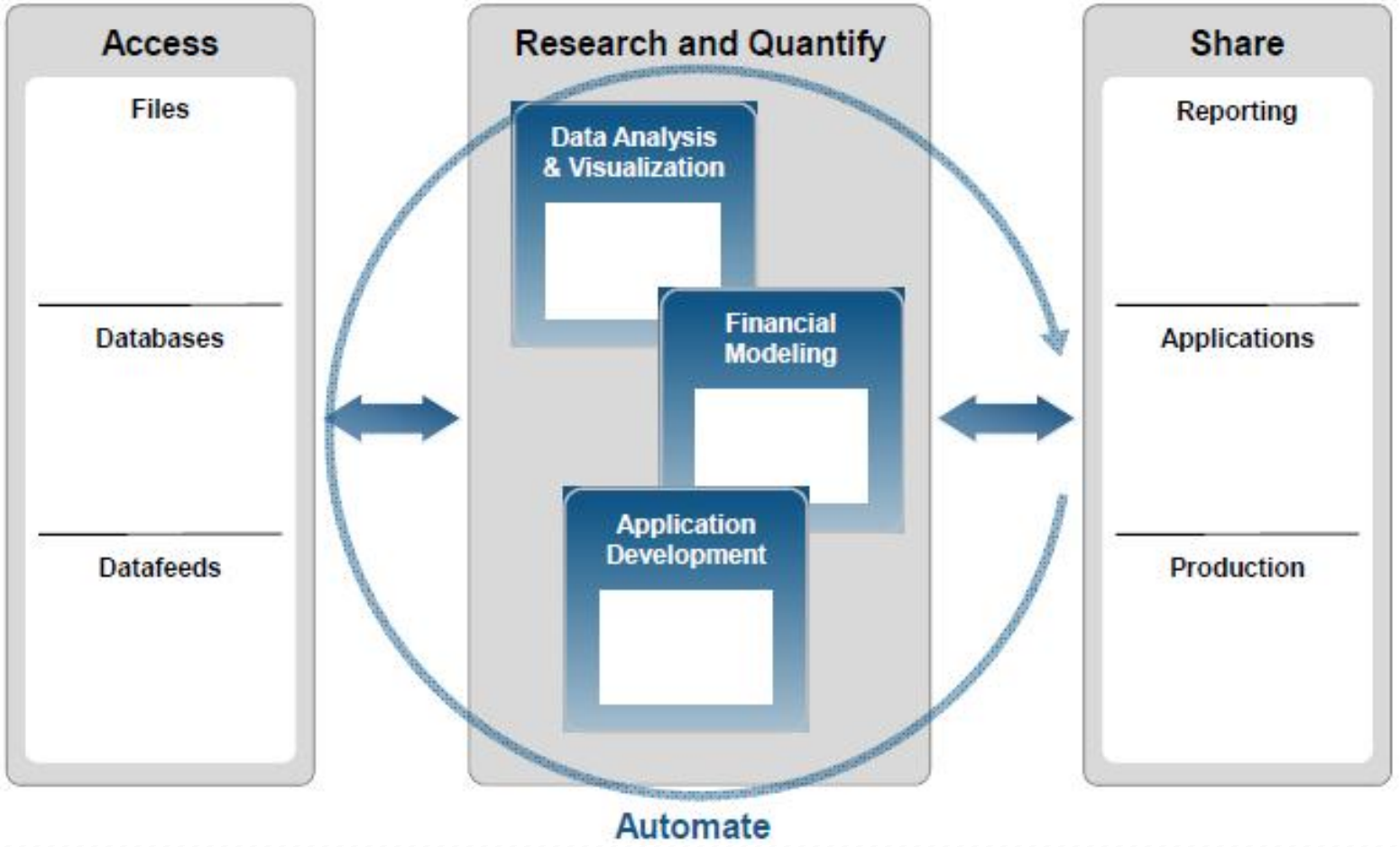
分析流程与思路

They need to find nuggets of truth in data and then explain it to the Business leaders" – Rchard Snee Emc

分析流程

流程图

分析是一种循环，一门艺术



分析思路

错误的模型得到的结果，不如不做！

思路

问：统计资料表明．大多数汽车事故出在中等速度的行驶中，极少的事故是出在大于150公里/小时的行驶速度上的。这是否就意味着高速行驶比较安全？

答：绝不是这样。统计关系往往不能表明因果关系。由于多数人是以中等速度开车，所以多数事故是出在中等速度的行驶中。

分析思路

思路

数据本身不会说谎！错的是你的解读！

问：统计数字表明，在亚利桑那州死于肺结核的人比其他州的人多。这是否就意味着亚利桑那州的气候容易生肺病？

答：正好相反。亚利桑那的气候对害肺病的人有好处，所以肺病患者纷纷前来，自然这就使这个州死于肺结核的平均数升高了。

分析思路

思路

多看多想，实战中累计经验！

问：有一个调查研究说脚大的孩子拼音比脚小的孩子好。这是否是说一个人脚的大小是他拼音能力的度量？

答：不是的。这个研究对象是一群年龄不等的孩子。它的结果实际上是因为年龄较大的孩子脚大些，他们当然比年幼的男子拼得好些。

分析思路

数据本身不会说谎！错的是你的解读！

思路

问：常常听说，汽车事故多数发生在离家不远的地方，这是否就意味着在离家很远的公路上行车要比在城里安全些呢？

答：不是，统计只不过反映了人们往往是在离家不远的地方开车，而很少在远处的公路上开车。

分析思路

数据分析挖掘是门科学，更是门艺术！

思路

问：有一项研究表明其一个国家的人民，喝牛奶和死于癌症的比例都很高。这是否说明是牛奶引起癌症呢？

答：不！这个国家老年人的比例也很高。由于癌症通常是年龄大的人易得，正是这个因素提高了这个国家癌症死亡者的比例。

个人观点

关于时间

真心讨厌很多人说自己忙却不知道为什么忙，在我看来，时间都是可以管理的，只有想做不想做，没有忙不忙一说。何况，很多人所谓的忙，也只是简单的重复而已。

关于软件

一直和朋友探讨一个问题：现在的各种软件越来越智能，越来越接近傻瓜化，导致了用户很多时间不必追究模型中具体的算法与应用条件，这到底是好是坏？软件的智能化不等于整个数据分析的智能化与自动化，如果没有对业务的深刻理解，建立的模型就没有说服力，更不用说使用模型进行决策了。

关于数据分析与挖掘

数据挖掘与统计分析区别：(1)DM处理大量数据更强势，无需太专业统计背景；(2)从大型数据库抓取所需数据并用专业软件角度来看，DM更符合企业需求；(3)纯就理论基础点来看，DM和统计分析在应用上存在差别，毕竟DM的目的是方便企业决策使用而非给统计学家检测用的。

关于使用

你会用啥就用啥，用什么顺手就用什么。记住三点：一：只要能达到目标的软件就是好软件；二：你研究的领域啥软件好用啥软件就是好软件 ；三：不要妄想用一个软件解决一切问题。

小测试

你觉得这个推论如何？

有人在微博上说：“男人比女人更耐寒！"。原因是最近一周淘宝销售女式羽绒服18万件，男士不到6万件。女式均单价358元，男装比女装均价略低10来块钱，你觉得这个结论怎么样？这样的推断错在哪里了？

1：女性消费者在淘宝上购物可能更多或者说更加热衷淘宝购物

2：在选择羽绒服这件事情上，男性可能更加喜欢实体店购物

3：不可忽略的是选择女士羽绒服中的用户多少是男性？是否有送情侣或者爱人的情况，这样就分析不出男人比女人更耐寒。

4：女性可能1个人买N件，不理性。

5：男人可以穿一件羽绒服穿好多年。。也不一定要年年买的。

6：羽绒服只是保暖的其中一个工具，还可能有其他保暖产品销售，最后取加权平均。

7：那些女装有可能是男人买的...

8：只是最近一周的而已 不具代表性

9：男装销量不如女装可能是由于男装的款式非常陈旧而女装的款式新颖，所以女装的销量比较大。

10：男人过冬不常穿羽绒服...穿风衣

第二部分

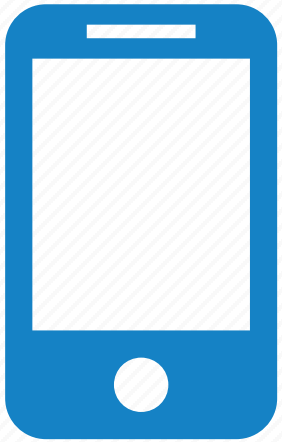
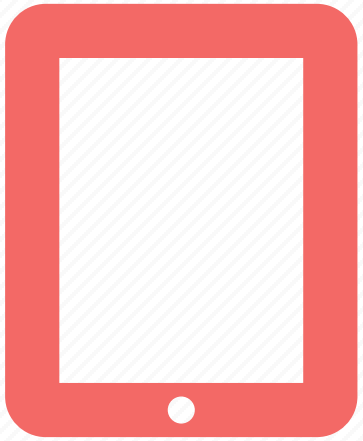
Second Part

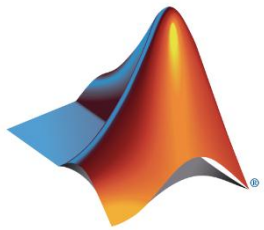
数据分析软件介绍

data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product" – Hillary Mason

软件对比

你看看你使用过哪些？

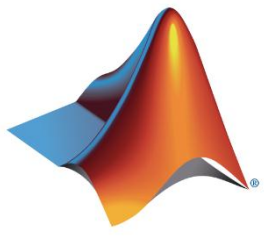
Python  VS  Others


软件对比

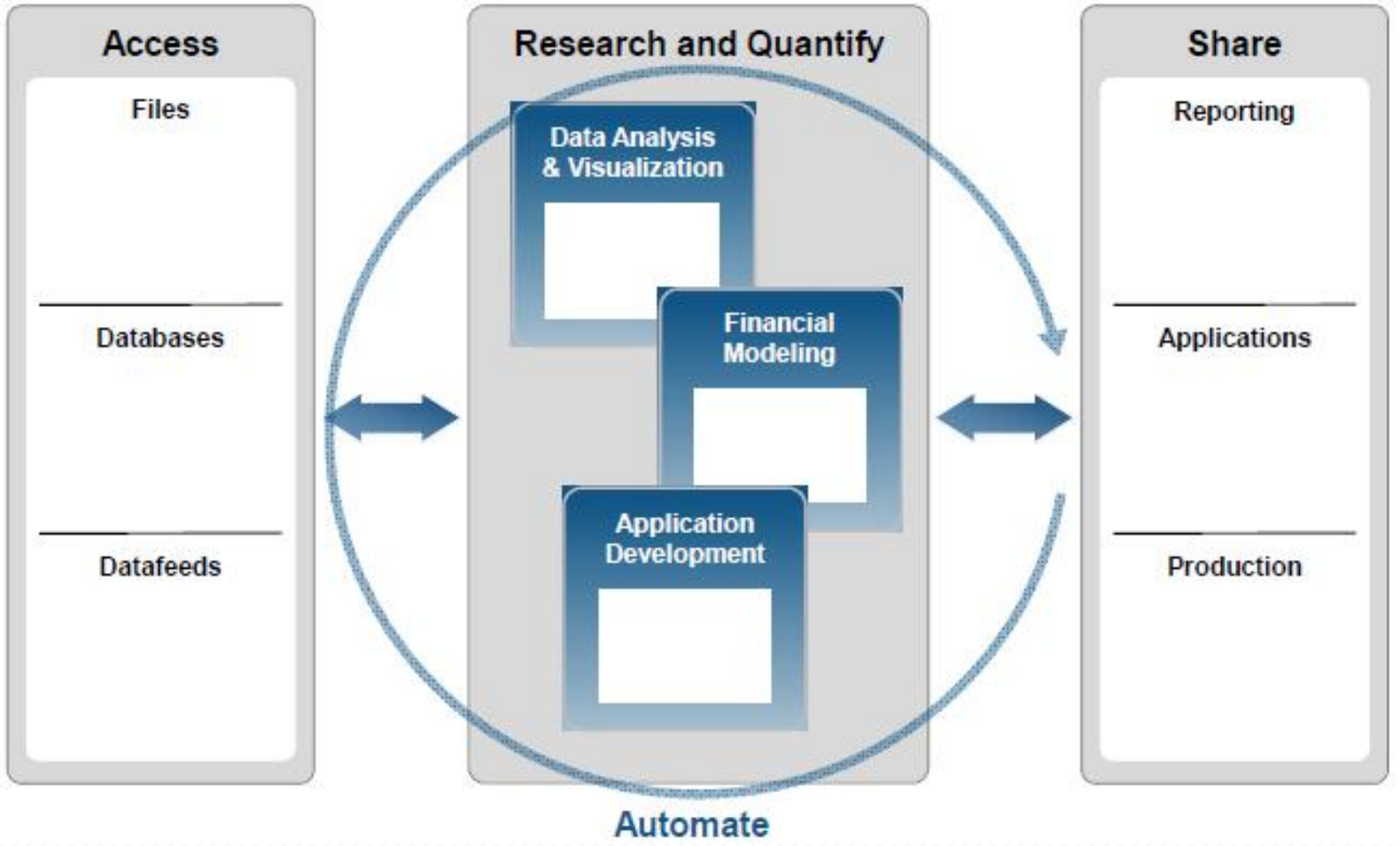
你看看你使用过哪些？

软件	收费	处理逻辑	版本更新	编程	用途
Python		*	***	***	***
MATLAB	*	*	**	**	***
R		*	***	***	*
SPSS	*	*	*	*	*
STATA	*	*	*	*	*
SAS	*	**	*	*	*
jmp	*	*	*	*	*
Excel	*	*	*	*	*



模型介绍

分析流程


Access
Files
Databases
Datafeeds
Research and Quantify
Data Analysis & Visualization
Financial Modeling
Application Development
Automate
Share
Reporting
Applications
Production

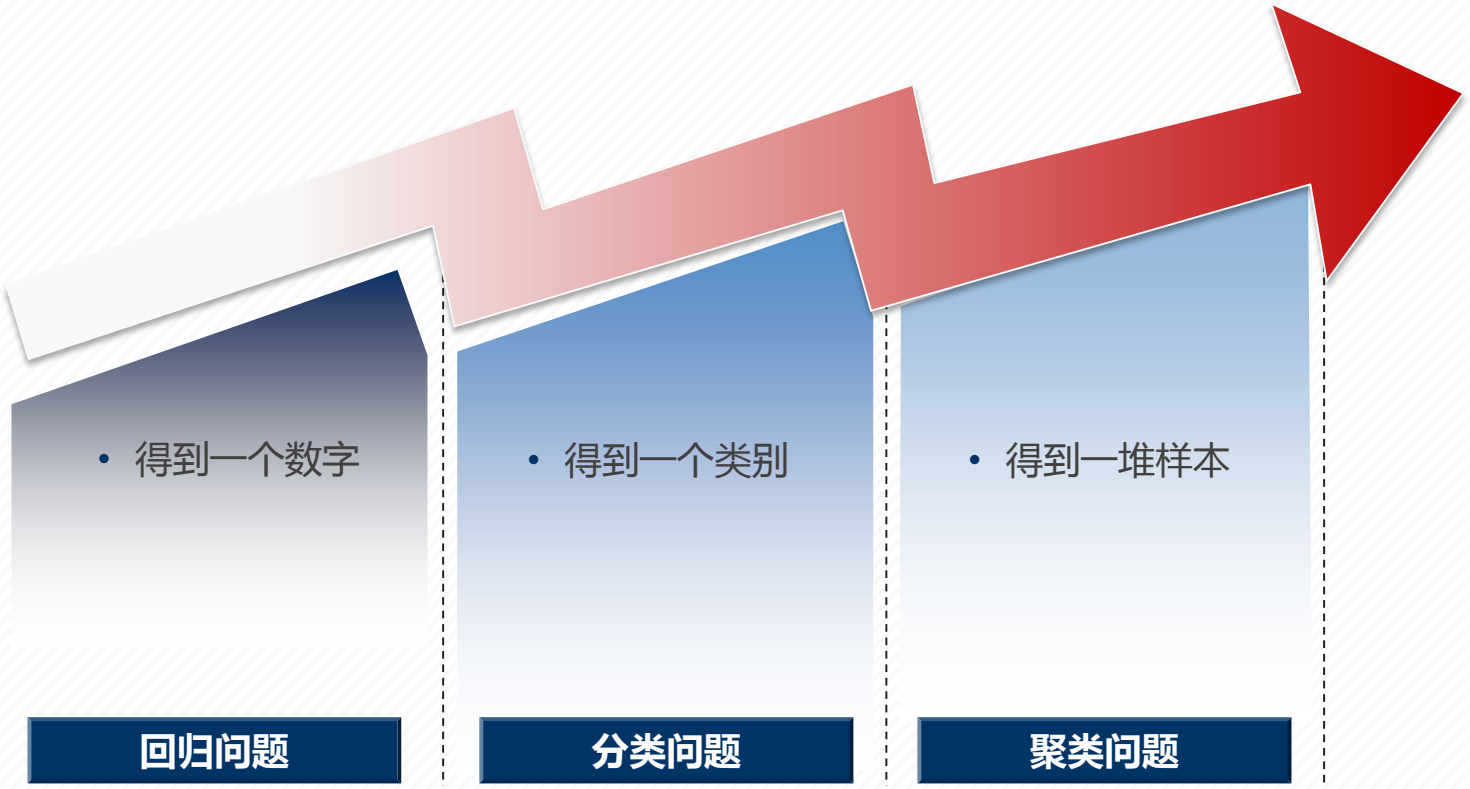
模型介绍

衡量标准

Generating a good model	accurate
	stable
	general
Ease of Use	generate a fit
	measure accuracy
	make predictions
	switch algorithm
	share results
Feature selection	uncorelated predictor
	corelated predictor

模型介绍

模型分类



模型介绍

模型分类

回归问题

模型

- 1 多元线性回归
- 2 多元非线性回归
- 3 广义线性回归模型
- 4 神经网络
- 5 曲线拟合

分类问题

模型

- 1 神经网络 2 逻辑回归
- 3 判别分析
- 4 朴素贝叶斯分类
- 5 SVM
- 6 决策树 7 组合算法

聚类问题

模型

- 1 K均值聚类
- 2 系统聚类
- 3 神经网络
- 4 模糊C均值聚类
- 5 高斯混合模型

神经网络

神经网络

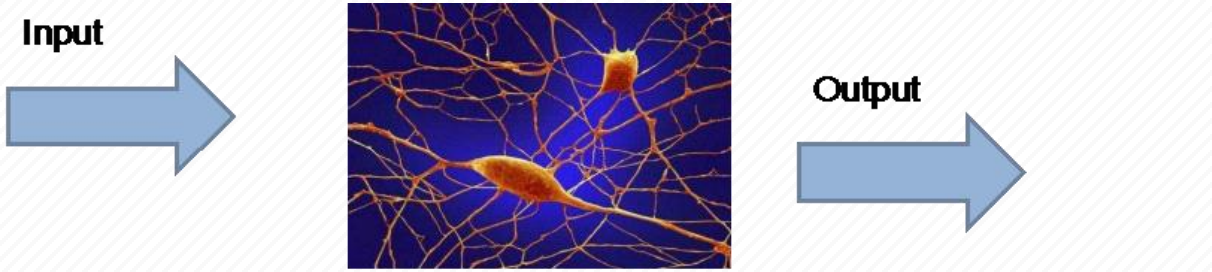
目的

1. 什么时候用神经网络？
2. 用神经网络分类还是回归？
3. Deterministic or Stochastic ?
4. 有监督？无监督？
5. Online or Offline ?
6. PC or other ?

数据挖掘从来都不是为了使用而使用，如果某些情况下不适合，不要大炮打蚊子！

神经网络

什么时候使用



1 不知道数学模型的时候

准确率=100%

准确率<100%

2 知道模型但是模型异常复杂

准确率=100%

准确率<100%

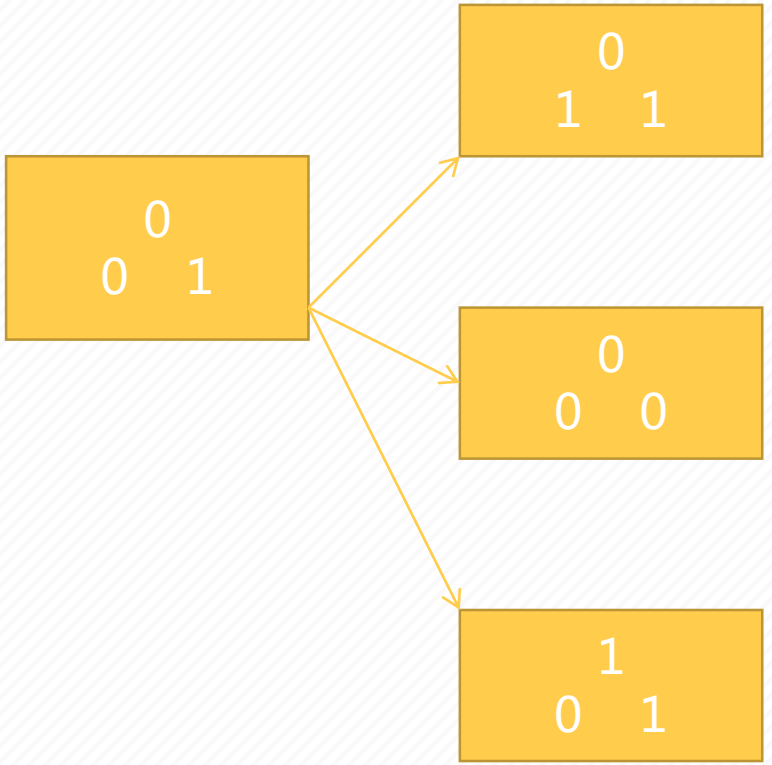
神经网络

分类还是回归？

分类	回归
BP	BP
SOM	RBF
PNN	ELMAN
….	…..

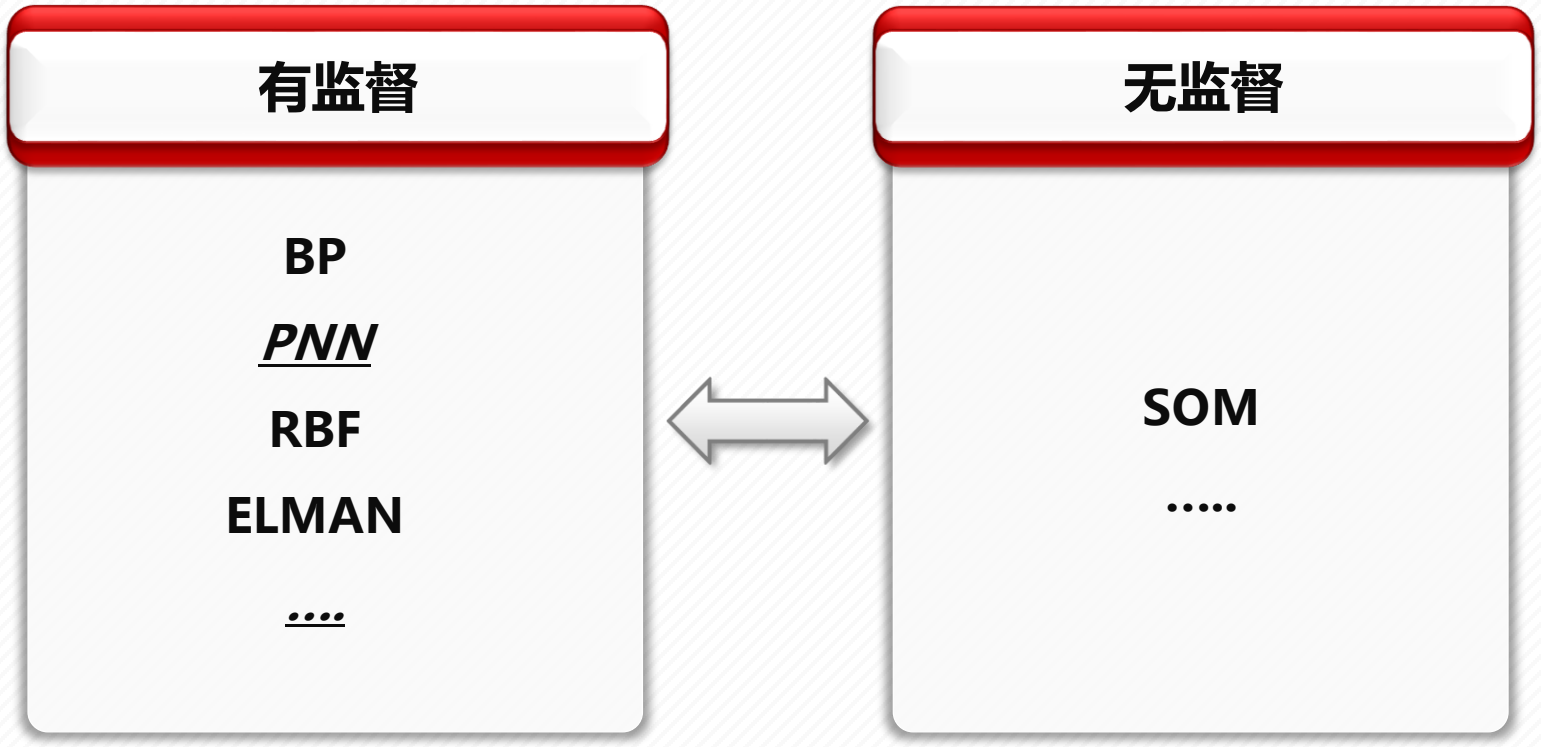
神经网络

改变概率的随机性



神经网络

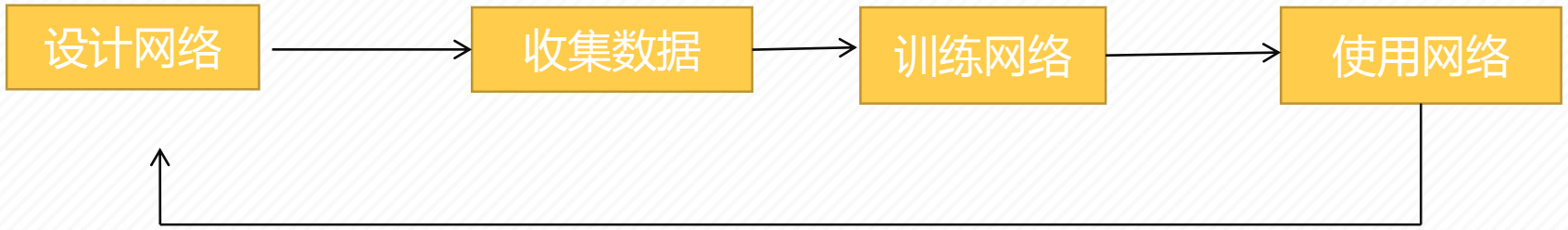
是否有监督



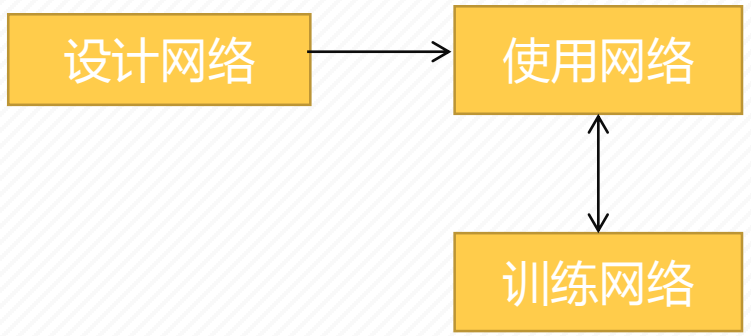
神经网络

Online or Offline ?

Offline:



Online:



神经网络

PC or Other ?

电脑：处理复杂网络

其他设备：尽量简化编程

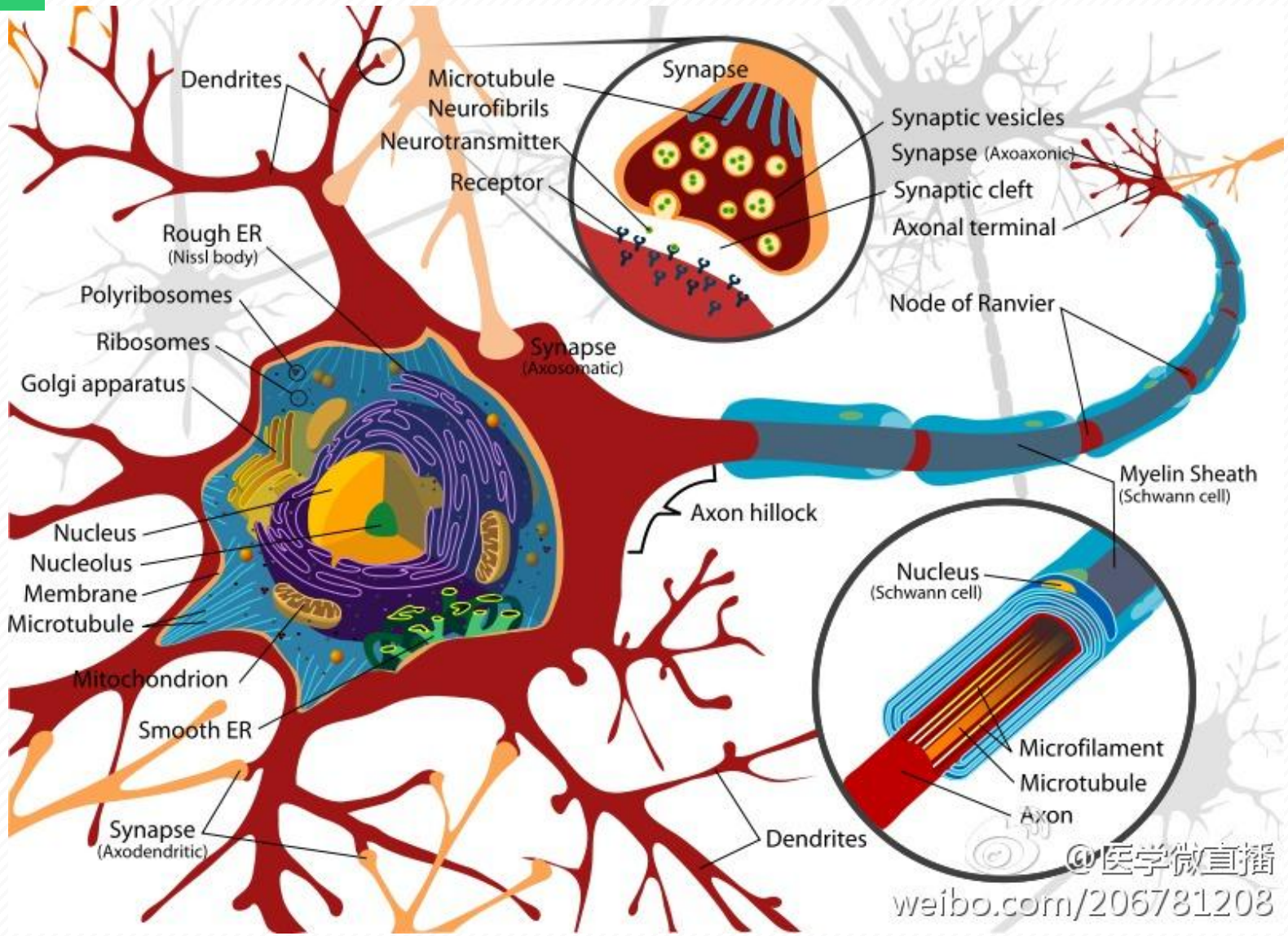
神经网络

ANN

1. 感知器
2. BP神经网络
3. RBF神经网络
4. SVM
5. Hopfield神经网络/Elman
6. SOM/kohonen

神经网络

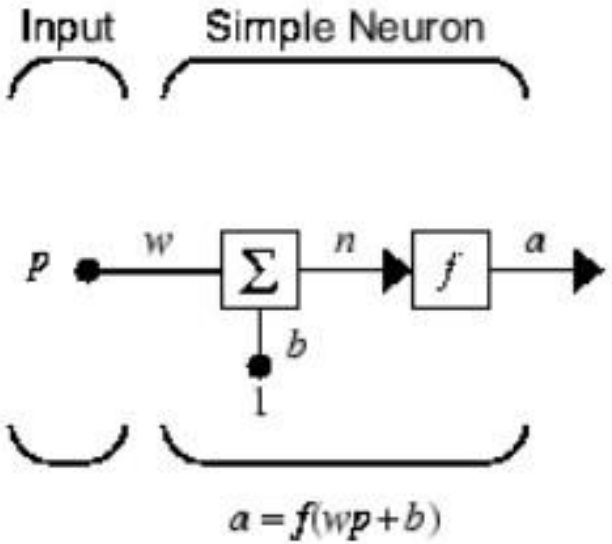
神经元


Dendrites
Microtubule
Neurofibrils
Neurotransmitter
Receptor
Synapse
Synaptic vesicles
Synapse (Axoaxonic)
Synaptic cleft
Axonal terminal
Rough ER
(Nissl body)
Polyribosomes
Ribosomes
Golgi apparatus
Synapse
(Axosomatic)
Node of Ranvier
Myelin Sheath
(Schwann cell)
Axon hillock
Nucleus
Nucleolus
Membrane
Microtubule
Mitochondrion
Smooth ER
Synapse
(Axodendritic)
Nucleus
(Schwann cell)
Microfilament
Microtubule
Axon
Dendrites
@医学微直播
weibo.com/206781208

神经元（单输入）

Simple Neuron

The fundamental building block for neural networks is the single-input neuron, such as this example.

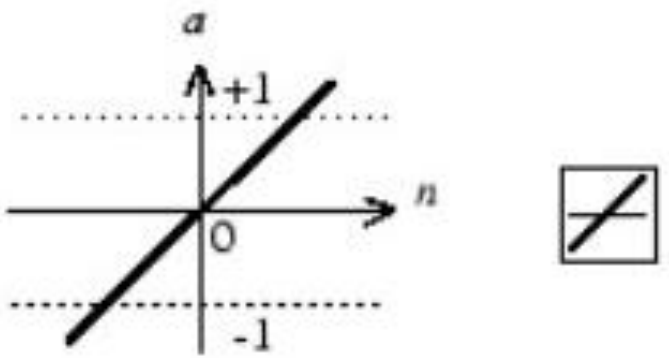


weight function (times/distance) **nnweight**
net function: (summation/multiplication) **nnnetinput**
transfer function：**nntransfer**

Note that w and b are both *adjustable* scalar parameters of the neuron. The central idea of neural networks is that such parameters can be adjusted so that the network exhibits some desired or interesting behavior. Thus, you can train the network to do a particular job by adjusting the weight or bias parameters.

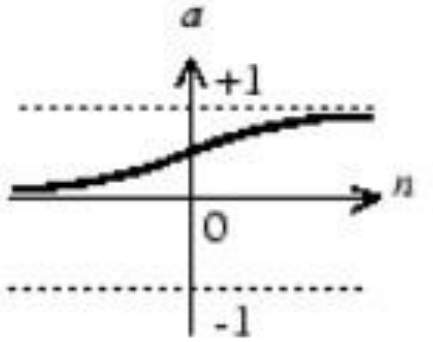
神经网络

传递函数



$a = purelin(n)$

Linear Transfer Function



$a = logsig(n)$

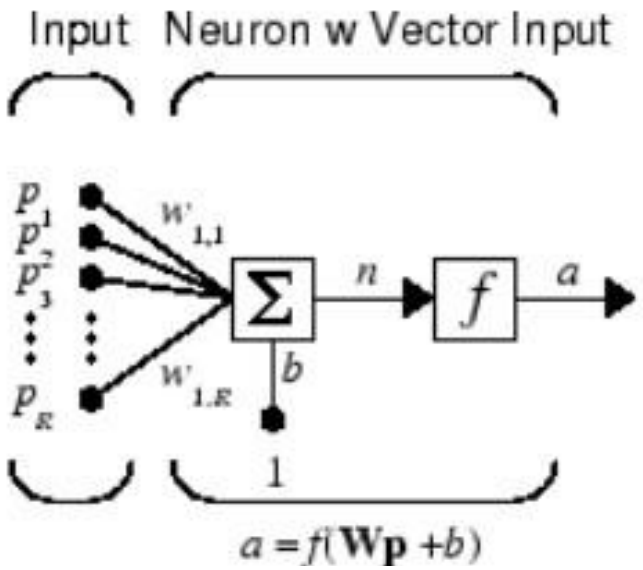
Log-Sigmoid Transfer Function

多层神经网络与BP神经网络：输出层传递函数

多层神经网络隐藏层传递函数----可微

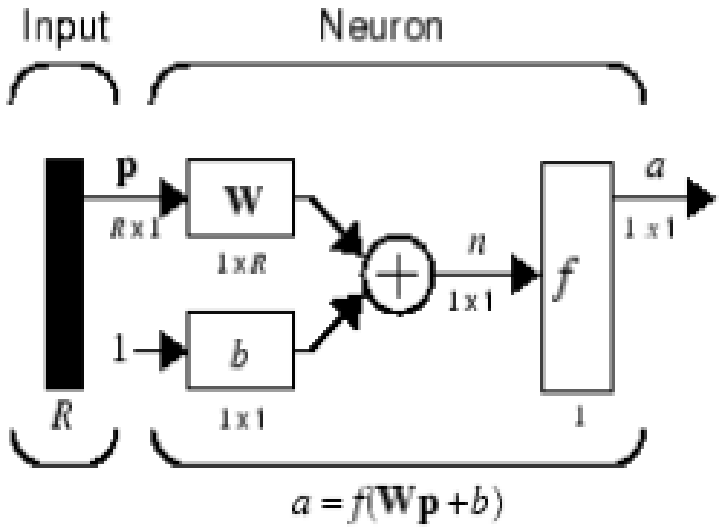
神经网络

神经元（多输入）



Where

R = number of elements in input vector



Where...

R = number of elements in input vector

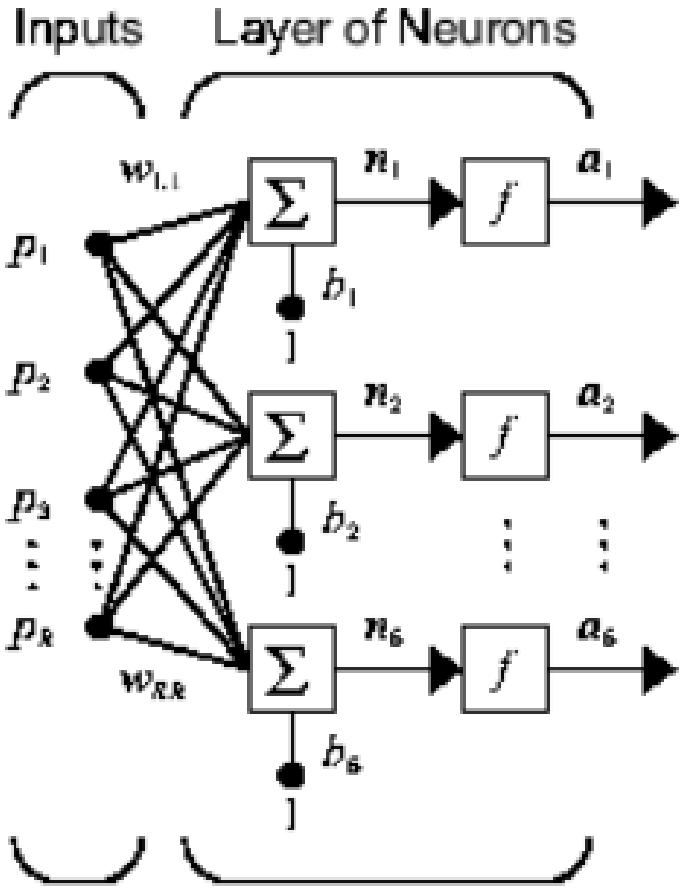
$$n = w_{1,1}p_1 + w_{1,2}p_2 + \ldots + w_{1,R}p_R + b$$

n = W*p + b

神经网络（层）

One Layer of Neurons

A one-layer network with R input elements and S neurons follows.



Where

R = number of elements in input vector

S = number of neurons in layer

神经网络

权重

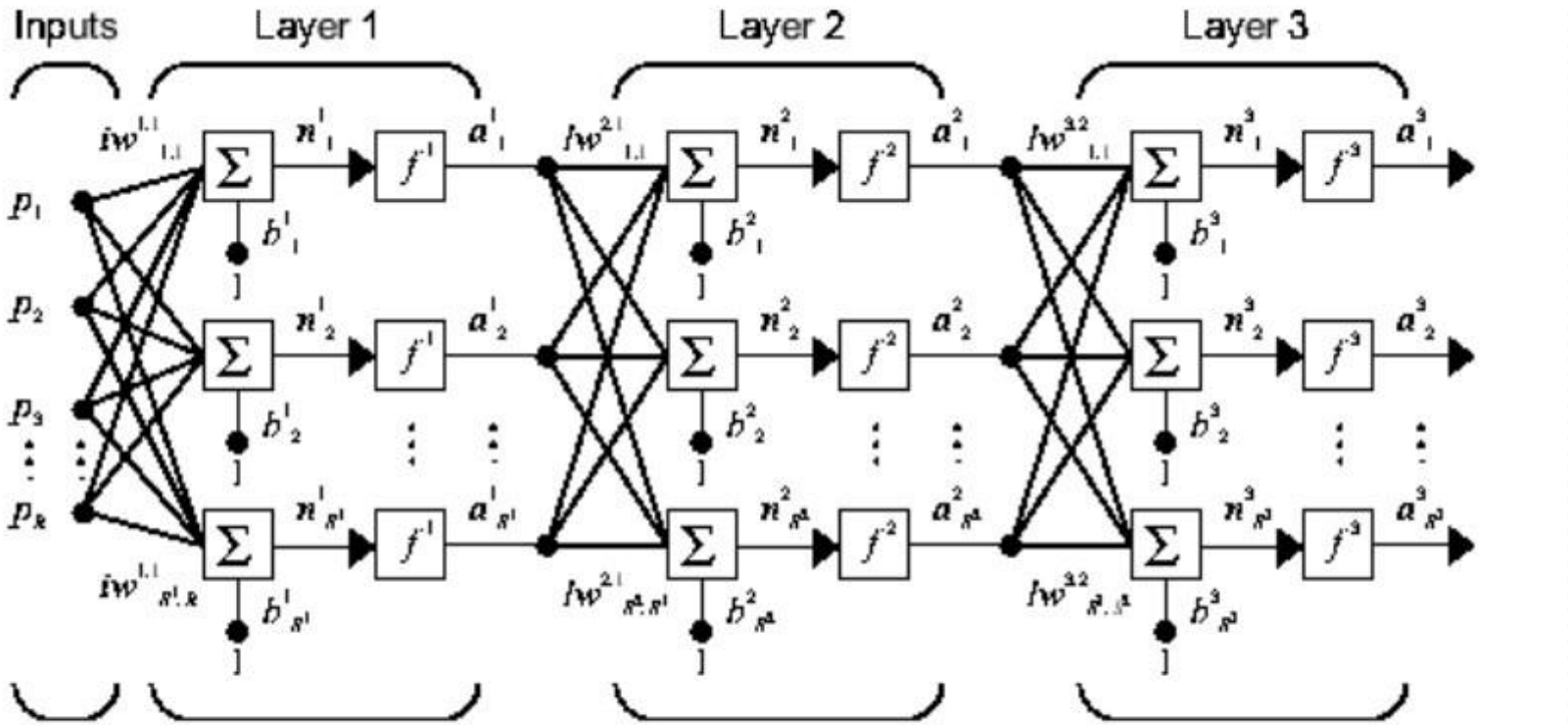
The input vector elements enter the network through the weight matrix **W**.

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\ & & & \\ w_{S,1} & w_{S,2} & \cdots & w_{S,R} \end{bmatrix}$$

Note that the row indices on the elements of matrix **W** indicate the destination neuron of the weight, and the column indices indicate which source is the input for that weight. Thus, the indices in $w_{1,2}$ say that the strength of the signal *from* the second input element *to* the first (and only) neuron is $w_{1,2}$.

神经网络

神经网络（多层）

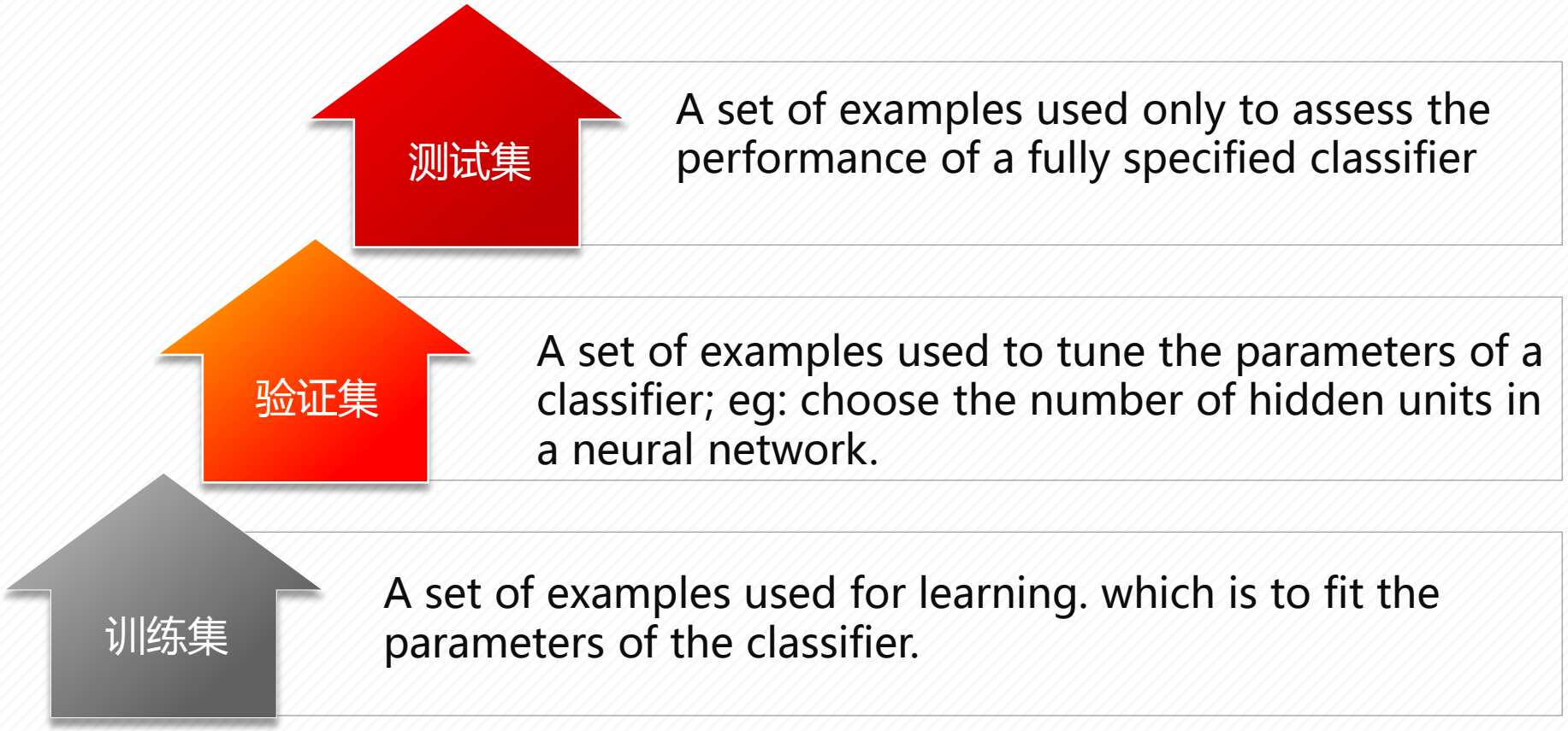


$$\mathbf{a}^1 = \mathbf{f}^1(\mathbf{IW}^{1,1}\mathbf{p}+\mathbf{b}^1) \qquad \mathbf{a}^2 = \mathbf{f}^2(\mathbf{LW}^{2,1}\mathbf{a}^1+\mathbf{b}^2) \qquad \mathbf{a}^3 = \mathbf{f}^3(\mathbf{LW}^{3,2}\mathbf{a}^2+\mathbf{b}^3)$$

$$\mathbf{a}^3 = \mathbf{f}^3(\mathbf{LW}^{3,2}\mathbf{f}^2(\mathbf{LW}^{2,1}\mathbf{f}^1(\mathbf{IW}^{1,1}\mathbf{p}+\mathbf{b}^1)+\mathbf{b}^2)+\mathbf{b}^3)$$

神经网络

数据集


测试集
A set of examples used only to assess the performance of a fully specified classifier
验证集
A set of examples used to tune the parameters of a classifier; eg: choose the number of hidden units in a neural network.
训练集
A set of examples used for learning. which is to fit the parameters of the classifier.


THANKS